

Centre de Physique Théorique* - CNRS - Luminy, Case 907
F-13288 Marseille Cedex 9 - France

AUTOMATIC BIASES CORRECTION

Roland TRIAY¹

Abstract

The key point limits to define the *statistical model* describing the data distribution. Hence, it turns out that the characteristics related to the so-called Inverse Tully-Fisher relation and the Direct relation are maximum likelihood (ML) estimators of different statistical models, and we obtain coherent distance estimates as long as the same model is used for the calibration of the TF relation and for the determination of distances. The choice of the model is motivated by reasons of *robustness* of statistics, which depends on selection effects in observation.

Key-Words : galaxies : distance scale, distances – statistical methods

October 1994
CPT-94/P.3082

anonymous ftp or gopher: cpt.univ-mrs.fr

*Unité Propre de Recherche 7061

¹and Université de Provence, Marseille

1. INTRODUCTION

The method of correcting biases in estimating the distances of galaxies is one of the major problem which must be solved for a better understanding of the cosmic velocity fields, see [3, 5, 7] and P.Teerikorpi (this conference). If one keeps in mind that any technique of fitting is intimately related to a *statistical model* [1] then one understands that the cause of the weak convergence of present debates, for arguing on the use of either the direct Tully-Fisher relation (DTF) or the inverse relation (ITF), interprets as an *unsufficiently handled formulation* of the problem. The obstacle toward a consensus can be overcome by arguing on the model instead of the technique of fitting. Most of the present contribution is a brief presentation of results obtained in [9].

2. BASICS OF THE BIASES CORRECTION

To ask oneself whether the statistical estimator (*statistic*) corresponds to the model parameter for which it has been made up, is indeed a sensible question. Generically, a statistic $\hat{\theta}$ of a given parameter θ provides us with an estimate

$$\hat{\theta}_N = \theta + \epsilon_N \tag{1}$$

within a (unkown) random error ϵ_N , where N denotes the sample size. Thus, the accuracy of such an estimate can be discussed only in terms of characteristics describing the probability law of ϵ_N . For example, it is clear that the smaller the variance of ϵ_N the more precise such an estimate, as long as it is not biased. By definition, “ $\hat{\theta}_N$ is biased when the *expected value* of ϵ_N is not zero”. While an unbiased statistic shows a smaller variance, it turns out that such a property is not essential, it can be reached asymptotically (i.e., for $N \rightarrow \infty$).

Actually, the typical problem of biases in the present fields of interest is intimately related to the question of whether the selection effects in observation are correctly taken into account in the statistical model. In other words, we easily understand that one can obtain unbiased statistics as long as the *probability density* (*pd*) describing the ϵ_N -distribution is known, which requires a “statistical modeling” of the data. At this point, which is the first step toward the understanding of any problem involving observations, nothing prevents us to use solely the *maximum likelihood* (ML) technique for obtaining suitable statistics. The enormous advantage of such an approach is to provide us unambiguously with a unique fitting technique, which prevents us from subjective speculations on diagrams.

2.1 The Statistical Model – The Method

The pd describing the distribution of observables reads

$$dP_{\text{obs}} = \frac{\phi}{P_{\text{th}}(\phi)} dP_{\text{th}}, \quad (2)$$

where $0 \leq \phi \leq 1$ is a *selection function* in observation, dP_{th} describes the distribution of intrinsic variables related to sources and $P_{\text{th}}(\phi) = \int \phi dP_{\text{th}}$ is the normalization factor. Obviously, *working hypotheses* are required in order to define the selection function ϕ (in term of observables) and the theoretical $pd dP_{\text{th}}$ (in term of intrinsic quantities). Hence, we can write the likelihood function[†] $\mathcal{L}_{\text{obs}} = \mathcal{L}_{\text{th}} - \ln(P_{\text{th}}(\phi))$, where \mathcal{L}_{th} corresponds to the $pd dP_{\text{th}}$, and the ML statistic is derived from the equation

$$\partial_{\theta} \mathcal{L}_{\text{obs}} = 0. \quad (3)$$

Note the feature which informs on the presence of biases : a θ -statistic related to equation $\partial_{\theta} \mathcal{L}_{\text{th}} = 0$ differs from $\hat{\theta}_N$ if $\partial_{\theta} P_{\text{th}}(\phi) \neq 0$.

If the sample is not peculiar then the ML statistic $\hat{\theta}_N$ provides us with the *most probable value* of θ within a given *accuracy*, although it is not necessarily unbiased. For recovering an accurate estimate, the ML statistic must be shifted by the expected value of ϵ_N ,

$$\theta \approx \hat{\theta}_N - P_{\text{obs}}(\epsilon_N), \quad (4)$$

while (in practice) such an approach might demand cumbersome calculations. However, according to the *Central limit theorem*, if N is large enough then one expects that the discrepancy is neglectable ($\epsilon_N \approx 0$), which means that the ML statistic is asymptotically unbiased. Finally, we easily understand that any result is warranted as long as the distribution of variables involved in the calculation is correctly described by such a model.

The calculation of the *mean absolute magnitude* of galaxies from a magnitude limited sample is a pedagogic example for comparing the ML approach to the Malmquist (1920) calculation [6]. The statistical model is based on – a Gaussian luminosity distribution function; – a uniform spatial distribution; – and a sharp cutoff at a limiting magnitude m_{lim} . Thus $dP_{\text{th}} \propto g_G(M; M_{\odot}, \sigma_M) dM e^{\beta \mu} d\mu$, where $\beta = 3 \ln 10 / 5$, and the selection function $\phi_m(m) = \theta(m_{\text{lim}} - (\mu + M))$, where θ denotes the Heaveside distribution function. Since the normalization factor $P_{\text{th}}(\phi_m) \propto \exp\left(\frac{\beta}{2} \sigma_M^2 - M_{\odot}\right)$ depends on M_{\odot} , the standard statistics are expected to be biased. Indeed, if σ_M is unknown then the ML equations provide us with the following system of unbiased

[†]Actually, it is more convenient to use its natural logarithm.

statistics

$$M_{\circ} = \langle M \rangle + \beta \sigma_M^2, \quad (5)$$

$$\sigma_M^2 = \frac{1}{2\beta^2} \left(\sqrt{1 + 4\beta^2 \langle (M - M_{\circ})^2 \rangle} - 1 \right), \quad (6)$$

which can be solved by Newton's method. Note that the ML approach generalizes the Malmquist (1920) solution.

3. ABOUT THE DISTANCE ESTIMATE OF GALAXIES

The goal is to estimate a distance modulus from the observed apparent magnitude $m = M + \mu$ and the distance estimator p , which gives a rough estimate of the absolute magnitude $M \approx a.p + b$ by means of the Tully-Fisher relation (for spirals) [10], or the Faber-Jackson relation (for ellipticals) [2]. The distribution of intrinsic quantities is described by $dP_{\text{th}} = \kappa(\mu)d\mu F(p, M)dpdM$, where $\kappa(\mu)$ accounts for the galaxies distribution in space and $F(p, M)$ for the distribution in the p - M plane[‡]. For reasons that become clear in the following, we describe the p - M distribution according to different statistical models

$$F(p, M)dpdM = g_G(\zeta; 0, \sigma_{\zeta})d\zeta \times \begin{cases} f_M(M; M_{\circ}, \sigma_M)dM & \text{(ITF)} \\ f_p(p; p_{\circ}, \sigma_p)dp & \text{(DTF)} \end{cases}, \quad (7)$$

where $\zeta = a.p + b - M$ accounts for the *intrinsic* dispersion about the TF-relation, it is assumed to be Gaussian distributed about zero and with standard deviation σ_{ζ} .

Table 1 gives the related ML statistics of parameters a , b and σ_{ζ} in term of statistics of the covariance (Cov), the standard deviation (Σ), the mean ($\langle . \rangle$) and the correlation coefficient (ρ). It is then clear that the identifications of a to the “slope” and b to the “zero-point” of the TF relation are model dependent. These statistics are valid as long as the working hypotheses (*Constraints*) are fulfilled, in particular the absence of p -selection effects. They must be corrected for a bias due to measurement errors, which also increase the dispersion. However, for typical samples, we obtain estimates with a relative (1σ) accuracy of 7% for a and 15% for b . The simulations show that the main source of error is actually due to the small size of the calibration sample (≈ 30 galaxies) instead of errors.

Hence, we understand that the choice of the model must be discussed as a *strategy*. Indeed, the ITF model is much less constraintfull than the DTF, which makes the related statistics more *robust* (see e.g. [4]). In the other hand,

[‡]It must be noted that this distribution is different from the one in the TF-diagram, which is described by a $pd \propto F(p, M)dpdM \int_{\mu} \phi\kappa(\mu)d\mu$.

Table 1: Calibration Statistics

	ITF	DTF
a	$\Sigma(M)^2/\text{Cov}(p, M)$	$\text{Cov}(p, M)/\Sigma(p)^2$
b	$\langle M \rangle - a\langle p \rangle$	$\langle M \rangle - a\langle p \rangle + \beta\sigma_\zeta^2$
σ_ζ	$ \rho(p, M) ^{-1}\Sigma(M)\sqrt{1 - \rho^2(p, M)}$	$\Sigma(M)\sqrt{1 - \rho^2(p, M)}$
<i>Constraints</i>	$\phi_p = 1$	$\begin{cases} \phi_p\phi_\mu = 1 \\ \phi_m(m) = \theta(m_{\text{lim}} - m) \\ \kappa(\mu) \propto \exp(\beta\mu) \\ f_p(p) = g_G(p; p_o, \sigma_p) \end{cases}$

one might expect that (in general) the more numerous the working hypotheses the more precise the related statistic, the simulations show that the accuracy increases of 5% in the DTF model. However, it is clear that if one of these hypotheses is not so correct then the estimate is bogus. In practice, such a characteristic forces us to prefer the ITF approach, because of the usual conditions in observation. Nevertheless, it turns out that both models show the same robustness if they are improved for taking into account p -selection effects (in prep.).

In order to estimate a likely distance modulus μ of a galaxy from the same statistical model we have to assume that the galaxy belongs to *the same population* of the calibration sample. According to the Bayesian schema[§], provided the observables $m = m_k$ and $p = p_k$, the distribution of possible outcomes reads $dP_{\text{obs}}(\mu \mid m_k, p_k) \propto \int_M \int_p \delta(m - m_k) \delta(p - p_k) dP_{\text{obs}}$, which gives

$$f_\mu(\mu; \mu_\circ^{(k)}, \sigma_\mu^{(k)}) d\mu \propto \kappa(\mu) g_G(\mu; \tilde{\mu}_k, \sigma_\zeta) d\mu \times \begin{cases} f_M(m_k - \mu; M_0, \sigma_M) & \text{(ITF)} \\ 1 & \text{(DTF)} \end{cases}$$

where $\tilde{\mu}_k = m_k - (a.p_k + b)$ is model dependent, the mean $\mu_\circ^{(k)}$ and the standard deviation $\sigma_\mu^{(k)}$ depend on working hypotheses which specify the functions κ and f_M . The value $\mu_\circ^{(k)}$ interprets as an unbiased estimate of the distance modulus. The difference between $\mu_\circ^{(k)}$ and $\tilde{\mu}_k$ is not a bias of Malmquist type but a *volume correction*, since the Dirac's distribution functions cancel the dependence of any selection function on m and on p . Finally, it is important to mention that if the distribution function f_μ is not symmetric about $\mu_\circ^{(k)}$ then this unbiased distance estimate does not necessarily correspond to the *most probable distance*

$$\check{\mu}_k = \tilde{\mu}_k + \sigma_\zeta^2 \partial_\mu \ln \kappa(\mu) + \sigma_\zeta^2 \begin{cases} \partial_\mu \ln f_M(m_k - \mu; M_0, \sigma_M) & \text{(ITF)} \\ 0 & \text{(DTF)} \end{cases}, \quad (8)$$

[§]It is preferred to the *frequentist* schema [4] because the sample has a unique element, μ interprets as a model parameter of the $pd \, dP_{\text{obs}}(m_k, p_k \mid \mu)$.

which is defined as the root of equation $\partial_\mu f_\mu(\mu; \mu_\circ^{(k)}, \sigma_\mu^{(k)}) = 0$. Therefore, we see that the problem of the distance estimate of individual galaxies depends on the choice of the “strategy of gambling” (i.e., either one minimizes the random error or one bets to the most likely value within a given accuracy). According to Eq. (8), it is important to note that the DTF statistic does not require information on the luminosity distribution function, which makes the related distance estimate more robust than the ITF one. Therefore, we understand that if p -selection effects are absent then it is more convenient to use the ITF model for the calibration step, while the DTF model is preferred for the distance estimate. The possibility to get benefit of both advantages is presented by S. Rauzy (this conference).

If $f_M = g_G$ and $\kappa(\mu) \propto e^{\beta\mu}$ then the distance estimates coincide,

$$\check{\mu}_k = \mu_\circ^{(k)} = \begin{cases} \frac{1}{1+\gamma^2} \left((\tilde{\mu}_k + \beta\sigma_\zeta^2) + \gamma^2(m_k - M_\circ) \right) & \text{(ITF)} \\ \tilde{\mu}_k + \beta\sigma_\zeta^2 & \text{(DTF)} \end{cases} \quad (9)$$

where $\gamma = \sigma_\zeta^{\text{ITF}}/\sigma_M$ is a tiny quantity. The formal comparison of statistics shows that the discrepancy is a random variable of zero mean and neglectable standard deviation. Moreover, if the estimation of the mean M_\circ limits to the calibration sample then both models provide us with *the same distance estimate*[¶].

References

- [1] Bigot G., Triay R., 1990, *Phys. Lett. A* **150**,236
- [2] Faber, S.M., Jackson, R. 1976, *ApJ* **204**,668.
- [3] Gouguenheim L., Bottinelli L., Fouqué P., Paturel G., Teerikorpi P., 1989, in *The quest for the Fundamental Constants in Cosmology*, XXIVth Moriond Astrophys. Meetings, eds J. Audouze and J. Tran Thanh Van, p.3
- [4] Hendry M.A., Simmons J.F.L. 1990, *Astron. Astrophys.* **237**,275
- [5] Lynden-Bell D., Faber S.M., Burstein D., 1988, *ApJ* **326**,19
- [6] Malmquist K., 1920, *Medd. Lund.* **22**,1
- [7] Teerikorpi P., 1990, *Astron. Astrophys.* **234**,1
- [8] Triay R., 1993, in *Cosmic Velocity Fields*, proceed. of the 9th Astrophysics Meeting IAP, Paris, eds. F.R. Bouchet & M. Lachièze-Rey.

[¶]Since we have the ML estimate $M_0 = a^{\text{ITF}}\langle p \rangle_1 + b^{\text{ITF}} + \beta(\Sigma_1(M))^2$.

- [9] Triay R., Lachèze-Rey M., Rauzy S., 1994, *Astron. Astrophys.* **289**, 19
- [10] Tully R.B., Fisher J.R. 1977, *Astron. Astrophys.* **54**, 661